How to model just about anything (but especially habitat)

EFB 390: Wildlife Ecology and Management

Dr. Elie Gurarie

September 28, 2023

Super fast primer on statistical modeling

Everything you need to know to do 95% of all wildlife modeling in less than an hour and **FOUR** (or **FIVE**) easy steps!!

I. Linear modeling

II. Multivariate modeling

III. Model selection

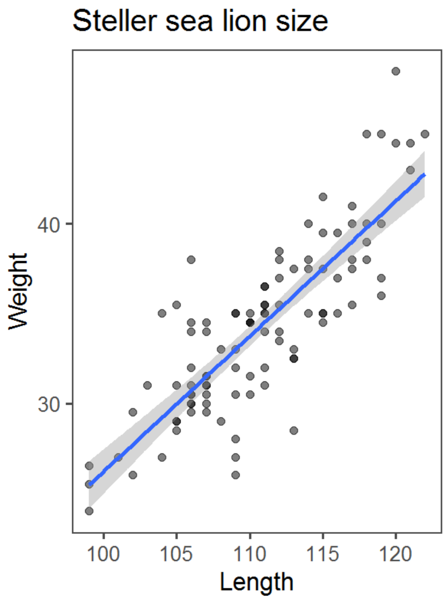
IV. Generalized linear modeling

- Poisson; Binomial

V. Prediction

Step I: Linear modeling

… is a very general method to quantifying relationships among variables.



X_i - is called:

- covariate
- independent variable
- explanatory variable

Y_i - is the property we are interested in modeling:

- response variable
- dependent variable

Note: There actually can be interest in wildlife studies to have models for **length** and **weight**, since **length** is easy to measure (e.g. from drones), but **weight** tells us more about physical condition and energetics.

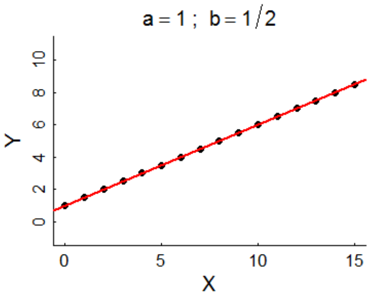
Steller sea lion (*Eumatopias jubatus*) pups.

Linear Models

Deterministic:

$$Y_i = a + bX_i$$

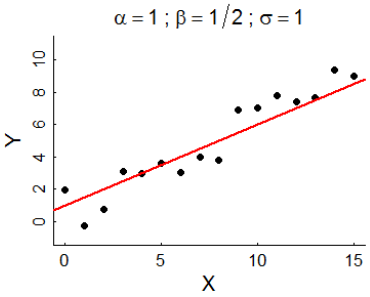
a - intercept; b - slope



Probabilistic:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

α - intercept; β - slope; ϵ - **randomness!**: $\epsilon_i \sim \mathcal{N}(0, \sigma)$



Fitting linear models is very easy in R!

Point Estimate

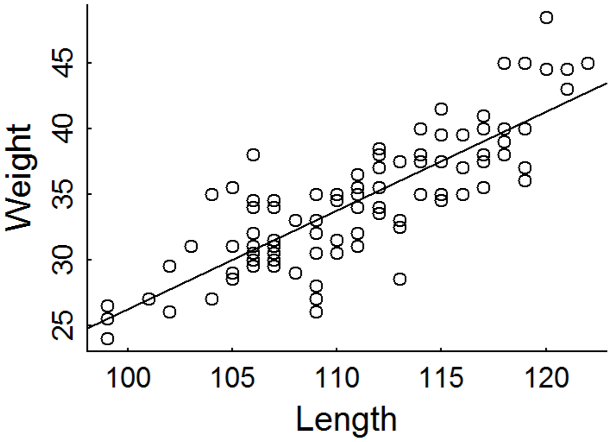
This command fits a model:

```
lm(Weight ~ Length, data = pups)
```

```
##
## Call:
## lm(formula = Weight ~ Length, data = pups)
##
## Coefficients:
## (Intercept)       Length
##    -49.1422       0.7535
```

So for **each 1 cm** of length, add another **754 grams**, i.e. `(\widehat{\beta} = 0.754)`

```
plot(Weight ~ Length, data = pups)
abline(my_model)
```



The `abline` puts a line, with intercept `a` and slope `b` onto a figure.

Some comments on linear models

$$Y_i \sim \alpha + \beta X_i + \epsilon_i$$

1. ϵ_i is **unexplained variation** or **residual variance**. It is often POORLY/WRONGLY referred to as "**error**". It is a **random variable**, NOT a **parameter**

2. A **better**, more sophisticated way to think of this model is not to focus on isolating the residual variance, but that the whole process is a random variable:

   $$Y_i \sim \mathcal{N}(\alpha + \beta X_i, \sigma)$$

   This is better because: (a) the three parameters (α, β, σ) are more clearly visible, (b) it can be "generalized". For example the **Normal distribution** can be a **Bernoulli distribution** (for binary data), or a **Poisson distribution** for count data, etc.

3. $\alpha + \beta X_i$ is the **predictor**, or the "modeled" portion. There can be any number of variables in the **predictor** and they can have different powers, so:

   $$Y_i \sim \mathcal{N}(\alpha + \beta X_i + \gamma Z_i + \delta X_i^2 + \nu X_i Z_i, \sigma)$$

   is also a **linear** model.

Statistical inference

Statistical inference is the *science / art* of observings *something* from a **portion of a population** and making statements about the **entire population**.

In practice - this is done by taking **data** and **estimating parameters** of a **model**. (This is also called *fitting* a model).

Two related goals:

1. obtaining a **point estimate** and a **confidence interval** (precision) of the parameter estimate.
2. Assessing whether particular (combinations of) factors, i.e. **models**, provide any **explanatory power**.

This is (almost always) done using **Maximum Likelihood Estimation**, i.e. an algorithm searches through possible values of the parameters that make the model **MOST LIKELY** (have the highest probability) given the data.

Another gratuitous sea lion picture.

Statistical output

~~~
```
##
## Call:
## lm(formula = Weight ~ Length, data = pups %>% subset(Island ==
##     "Raykoke"))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.498 -1.718  0.023  1.764  7.276
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -49.14222    5.75796  -8.535 1.81e-13 ***
## Length        0.75345    0.05193  14.510  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.761 on 98 degrees of freedom
## Multiple R-squared:  0.6824,    Adjusted R-squared:  0.6791
## F-statistic: 210.5 on 1 and 98 DF,  p-value: < 2.2e-16
```
~~~

1. Point estimates and confidence intervals

Intercept (α): -49.14 ± 11.5

Slope (β): 0.75 ± 0.104

2. Is the model a good one?

p-values are very very small, in particular for **slope**

Proportion of variance explained is high:

$$R^2 = 0.68$$

Models and Hypotheses

Every *p*-value is a Hypothesis test.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-49.142	5.758	-8.535	0
Length	0.753	0.052	14.510	0

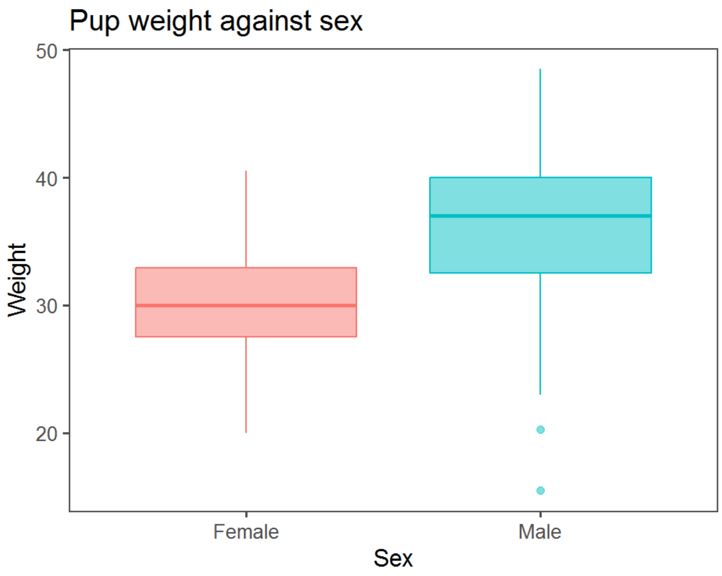
- First hypothesis test: H_0 intercept = 0
- Second hypothesis: H_0 slope = 0

Both null-hypotheses strongly rejected.

Linear modeling with a discrete factor

$$Y_{ij} = \alpha + \beta_i + \epsilon_{ij}$$

i is the index of sex (*Male* or *Female*), so there are two "Sex effects" - β_1 and β_2 representing the effect of the sex group; j is the index of the individual within each sex group i.



```
lm(Weight ~ Sex, data = pups)
```

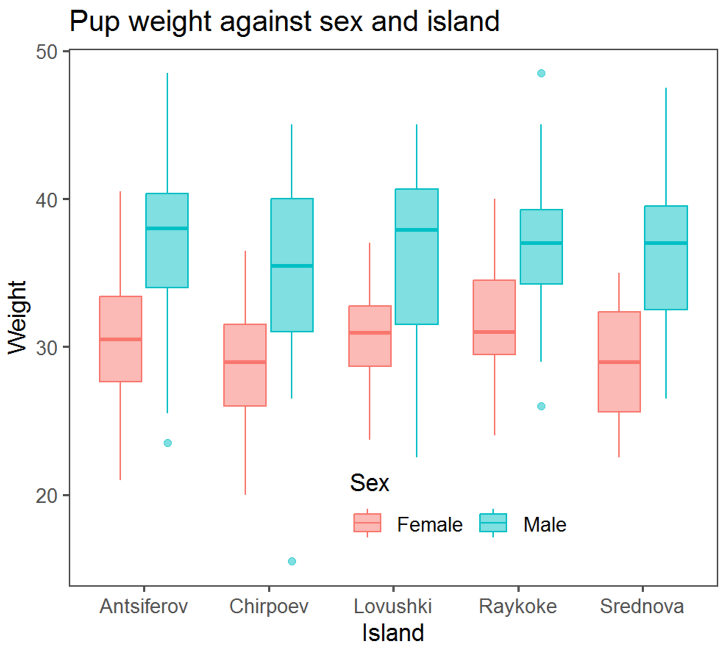
term	estimate	std.error	statistic	p.value
(Intercept)	30.151	0.317	95.119	<2e-16
SexMale	6.149	0.429	14.337	<2e-16

Intercept here means mean **female** weight.

Note - this is very similar to a *t*-test comparing two means (baby stats).

Linear modeling with multiple factors

Very easy to extend this to more complicated models!



$$Y_{ijk} = \alpha + \beta_i \, \text{Island}_{ijk} + \gamma_j \, \text{Sex}_{ijk} + \epsilon_{ijk}$$

```
lm(Weight ~ Island + Sex, data = pups)
```

term	estimate	std.error	statistic	p.value
(Intercept)	31.04	0.54	57.62	<1e-16
IslandChirpoev	-2.23	0.67	-3.34	0.001
IslandLovushki	-0.84	0.67	-1.26	0.21
IslandRaykoke	0.14	0.67	0.21	0.83
IslandSrednova	-1.50	0.67	-2.24	0.03
SexMale	6.14	0.42	14.47	1e-16

Analysis of Variance (ANOVA)

Is a technique for seeing which effect in a model is **significant**. Each row tests a **hypothesis** that the effect coefficients are non-zero.

In this model, we include an **interaction**, asking: "*Do different Islands have different patterns among Sexes? (and vice versa)*"

```
lm(Weight ~ Island * Sex, data = pups)
```

```
Analysis of Variance Table

Response: Weight
             Df  Sum Sq Mean Sq  F value    Pr(>F)
Island        4   443.3   110.8   5.0114 0.0005763 ***
Sex           1  4623.9  4623.9 209.0758 < 2.2e-16 ***
Island:Sex    4    71.4    17.9   0.8075 0.5207439
Residuals   488 10792.6    22.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

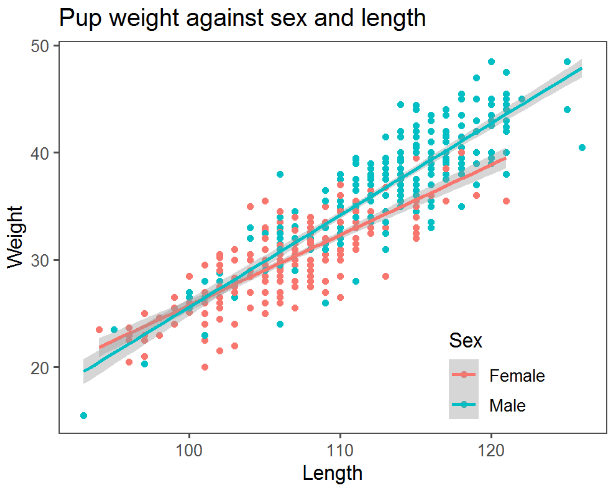
Non-significant interaction term

Interpretation:

- Differences between SEXES very significant (very very small p-value)
- Differences among ISLANDS very significant (small p-value)
- SEX differences among ISLANDS consistent (large **interaction** *p*-value)
- ISLANDS differences between SEXES consistent (large **interaction** *p*-value)

Combining **continuous** and **categorical** variables

Exploratory plot



It looks like, maybe, there are different body proportions for **MALES** and **FEMALES**.

ANOVA table confirms our suspicion!

```
Analysis of Variance Table

Response: Weight
            Df  Sum Sq Mean Sq  F value    Pr(>F)
Length       1 12413.8 12413.8 1957.969 < 2.2e-16 ***
Sex          1   257.3   257.3   40.582 4.321e-10 ***
Length:Sex   1   128.1   128.1   20.208 8.662e-06 ***
Residuals  494  3132.0     6.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Highly significant **interaction** term.

Step III: Model Selection

ANOVA is helpful for "nested" models, where each one is a subset of another more complex one. For comparing a **set of competing,non-nested** models, we use .

ΔAIC table

	Model	k	R2	logLik	AIC	dAIC
M0	1	1	0.000	-1569.5	3143.1	835.8
M1	Island	5	0.028	-1562.5	3137.0	829.7
M2	Sex	2	0.293	-1483.2	2972.4	665.1
M3	Length	2	0.779	-1193.4	2392.8	85.5
M4	Length + Sex	3	0.795	-1174.5	2357.0	49.7
M6	Length * Sex	4	0.803	-1164.5	2339.0	31.7
M5	Length + Sex + Island	7	0.811	-1155.0	2325.9	18.6
M7	**Length * Sex + Island**	**8**	**0.818**	**-1144.6**	**2307.3**	**0.0**
M8	Length * Sex * Island	20	0.824	-1137.1	2316.1	8.8

Degrees of freedom *k*:

- Number of estimated parameters. Measure of *complexity*.

Coefficient of determination R^2:

- Percent variation explained. It ALWAYS increases the more complex the model.
- Is is always zero for the **NULL** model.

log-likelihood $\log(\mathcal{L})$**:**

- Total probability score of model. It ALWAYS increases the more complex the model.

Akaike Information Criterion:

- $AIC = -2\log(\mathcal{L}) + 2\,k$
- A measure of model quality.
- Smaller is better It starts getting bigger if the model complexity gets too high.
- **The lowest AIC value is the "best" model.**
- (but within 2 ΔAIC is pretty much equivalent to best)

AIC in action: **What predicts ungulate body size?**

Quality (Nitrogen)? or Type (browse/grass)?

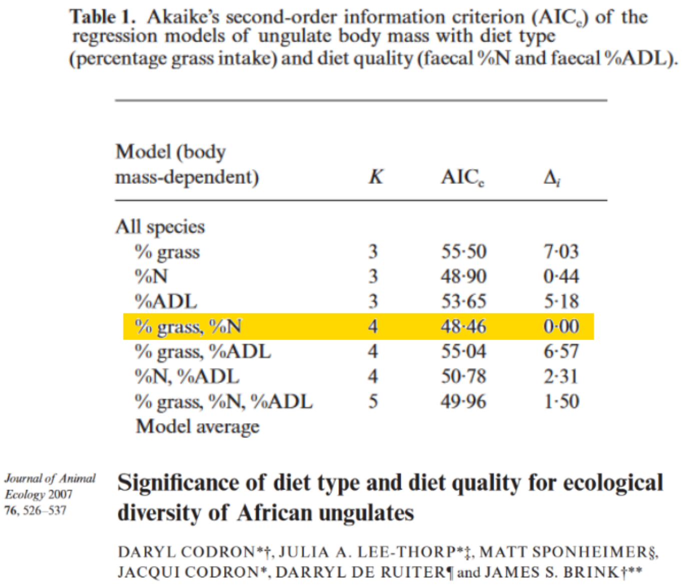


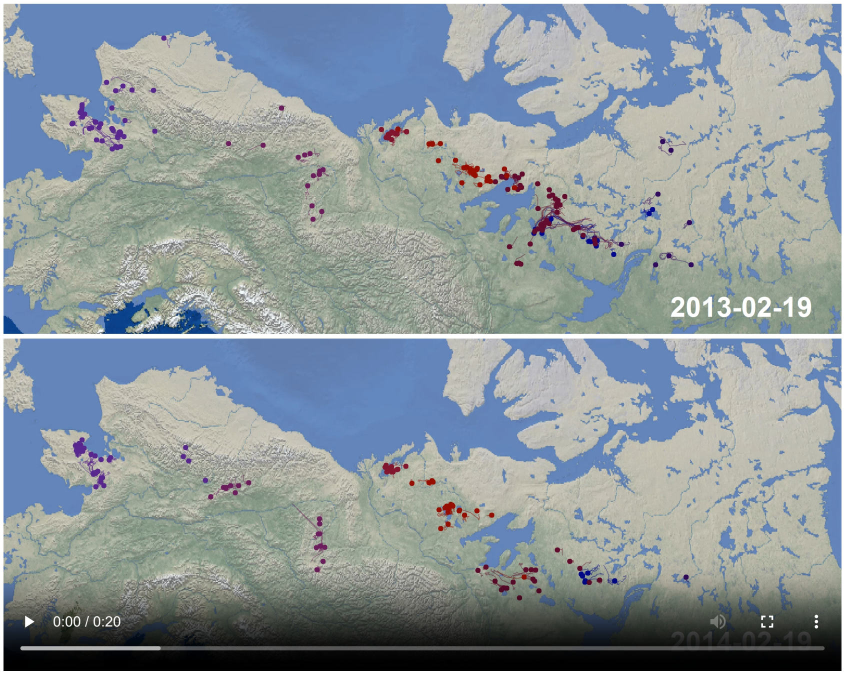
Table 1. Akaike's second-order information criterion (AIC$_c$) of the regression models of ungulate body mass with diet type (percentage grass intake) and diet quality (faecal %N and faecal %ADL).

Model (body mass-dependent)	K	AIC$_c$	Δ_i
All species			
% grass	3	55·50	7·03
%N	3	48·90	0·44
%ADL	3	53·65	5·18
% grass, %N	4	48·46	0·00
% grass, %ADL	4	55·04	6·57
%N, %ADL	4	50·78	2·31
% grass, %N, %ADL	5	49·96	1·50
Model average			

Journal of Animal Ecology 2007
76, 526–537

Significance of diet type and diet quality for ecological diversity of African ungulates

DARYL CODRON*†, JULIA A. LEE-THORP*‡, MATT SPONHEIMER§, JACQUI CODRON*, DARRYL DE RUITER¶ and JAMES S. BRINK†**

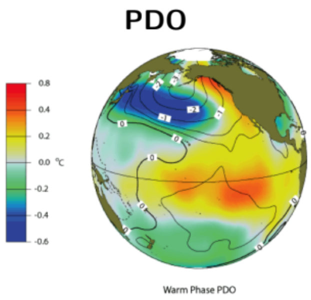


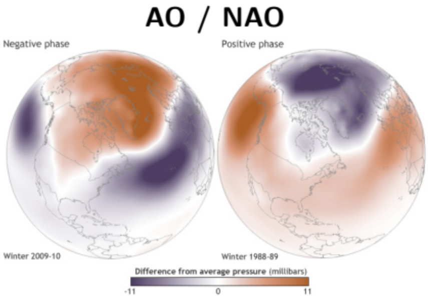
Caribou spring migrations

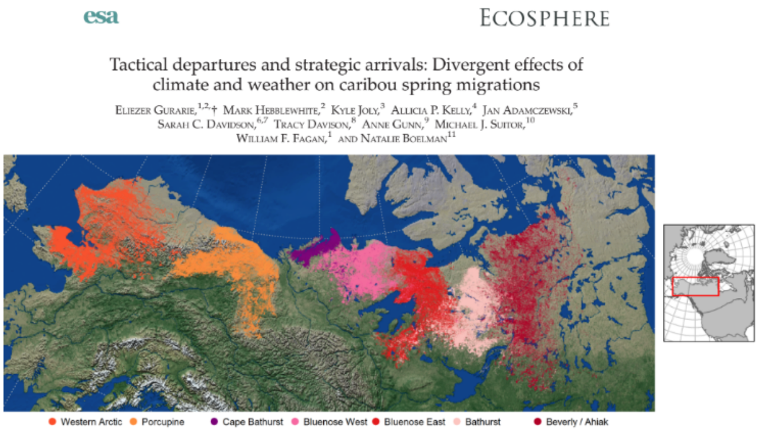
Remarkable temporal synchrony at a continental scale.

Could the synchrony be driven by global weather drivers?

Pacific Decadal Oscillation, Arctic Oscillation, North Atlantic Oscillation: determine whether the winter is wet & snowy or dry & cold.







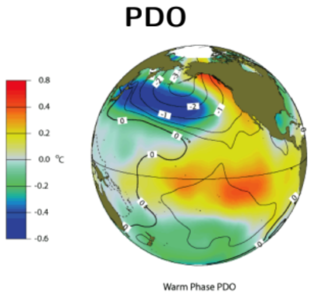
esa ECOSPHERE

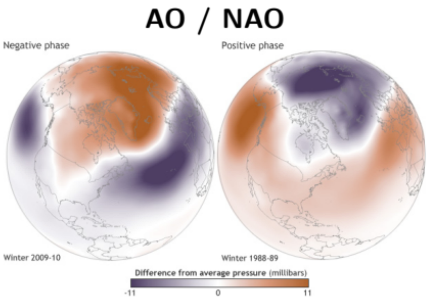
Tactical departures and strategic arrivals: Divergent effects of climate and weather on caribou spring migrations

Eliezer Gurarie,[1,2,†] Mark Hebblewhite,[2] Kyle Joly,[3] Allicia P. Kelly,[4] Jan Adamczewski,[5] Sarah C. Davidson,[6,7] Tracy Davison,[8] Anne Gunn,[9] Michael J. Suitor,[10] William F. Fagan,[1] and Natalie Boelman[11]

ΔAIC Table 1: **Departure time**

... driven by LARGE climate oscillations.





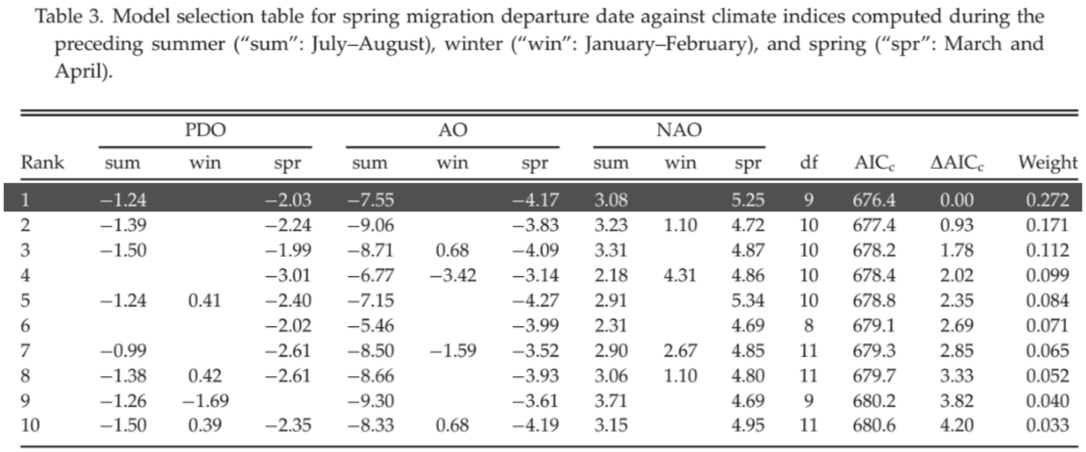
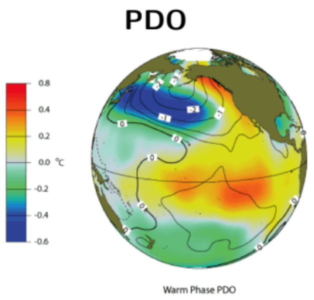


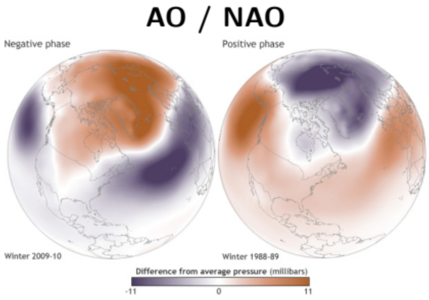
Table 3. Model selection table for spring migration departure date against climate indices computed during the preceding summer ("sum": July–August), winter ("win": January–February), and spring ("spr": March and April).

	PDO			AO			NAO						
Rank	sum	win	spr	sum	win	spr	sum	win	spr	df	AIC$_c$	ΔAIC$_c$	Weight
1	−1.24		−2.03	−7.55		−4.17	3.08		5.25	9	676.4	0.00	0.272
2	−1.39		−2.24	−9.06		−3.83	3.23	1.10	4.72	10	677.4	0.93	0.171
3	−1.50		−1.99	−8.71	0.68	−4.09	3.31		4.87	10	678.2	1.78	0.112
4			−3.01	−6.77	−3.42	−3.14	2.18	4.31	4.86	10	678.4	2.02	0.099
5	−1.24	0.41	−2.40	−7.15		−4.27	2.91		5.34	10	678.8	2.35	0.084
6			−2.02	−5.46		−3.99	2.31		4.69	8	679.1	2.69	0.071
7	−0.99		−2.61	−8.50	−1.59	−3.52	2.90	2.67	4.85	11	679.3	2.85	0.065
8	−1.38	0.42	−2.61	−8.66		−3.93	3.06	1.10	4.80	11	679.7	3.33	0.052
9	−1.26	−1.69		−9.30		−3.61	3.71		4.69	9	680.2	3.82	0.040
10	−1.50	0.39	−2.35	−8.33	0.68	−4.19	3.15		4.95	11	680.6	4.20	0.033

ΔAIC Table 2: **Arrival time**

... completely independent of climate!





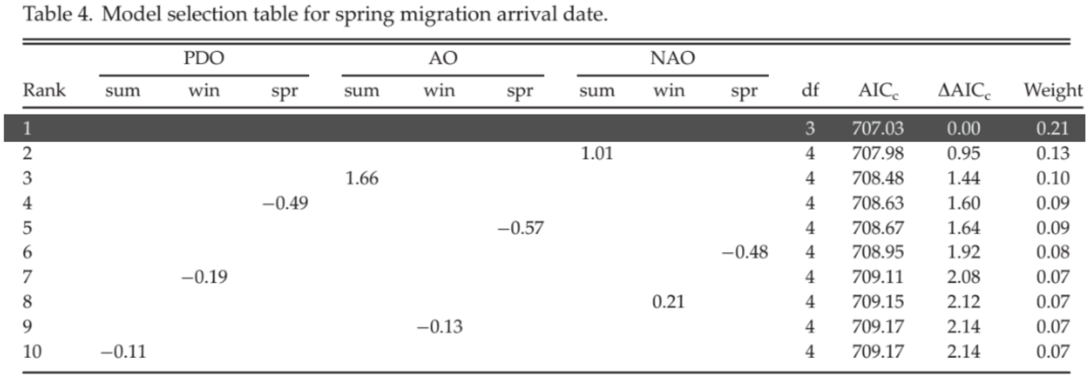


Table 4. Model selection table for spring migration arrival date.

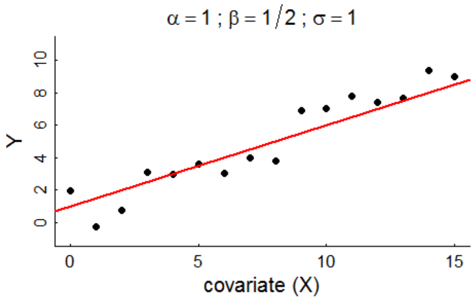
	PDO			AO			NAO						
Rank	sum	win	spr	sum	win	spr	sum	win	spr	df	AIC_c	ΔAIC_c	Weight
1										3	707.03	0.00	0.21
2							1.01			4	707.98	0.95	0.13
3				1.66						4	708.48	1.44	0.10
4			−0.49							4	708.63	1.60	0.09
5						−0.57				4	708.67	1.64	0.09
6									−0.48	4	708.95	1.92	0.08
7		−0.19								4	709.11	2.08	0.07
8								0.21		4	709.15	2.12	0.07
9					−0.13					4	709.17	2.14	0.07
10	−0.11									4	709.17	2.14	0.07

Step IV: *Generalized* linear modeling

Normal Model

$$Y_i \sim \mathcal{N}ormal(\alpha_0 + \beta_1 X_i, \sigma)$$

Models continuous data with a "normal-like" distribution.

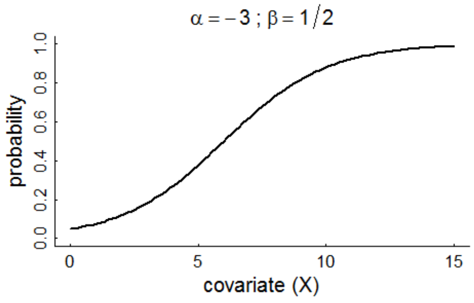


Binomial model

$$Y_i \sim \mathcal{B}ernoulli\left(\frac{\exp(\alpha + \beta X_i)}{1 + \exp(\alpha + \beta X_i)}\right)$$

There's some *probability* of something happening that depends on the predictor X.

Bernoulli just means the data are all 0 or 1.



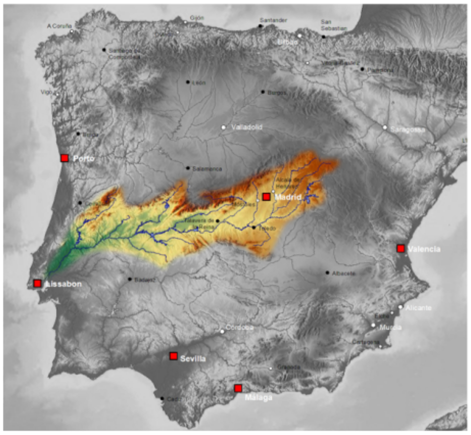
This models **presence/absence**, **dead/alive**, **male/female** other response variables with **2** possible outcomes.

What factors predict occurence of *Solea solea* larvae?

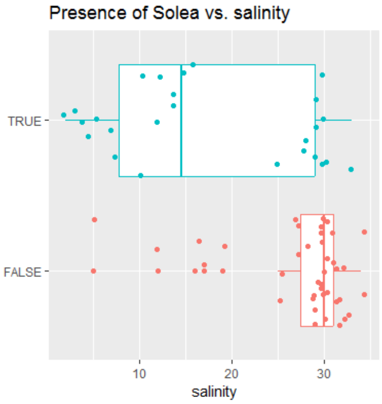
Sampled in the estuary of the Tejo river in Portugal

- Lots of environmental factors in data

depth	temp	salinity	transp	gravel	large_sand	fine_sand	mud	presence
3.0	20	30	15	3.74	13.15	11.93	71.18	**0**
2.6	18	29	15	1.94	4.99	5.43	87.63	**0**
2.6	19	30	15	2.88	8.98	16.85	71.29	**1**
2.1	20	29	15	11.06	11.96	21.95	55.03	**0**
3.2	20	30	15	9.87	28.60	19.49	42.04	**0**
3.5	20	32	7	32.45	7.39	9.43	50.72	**0**



Presence of *Solea solea* against **salinity**



Modeling is EXACTLY the same as **linear regression** except:

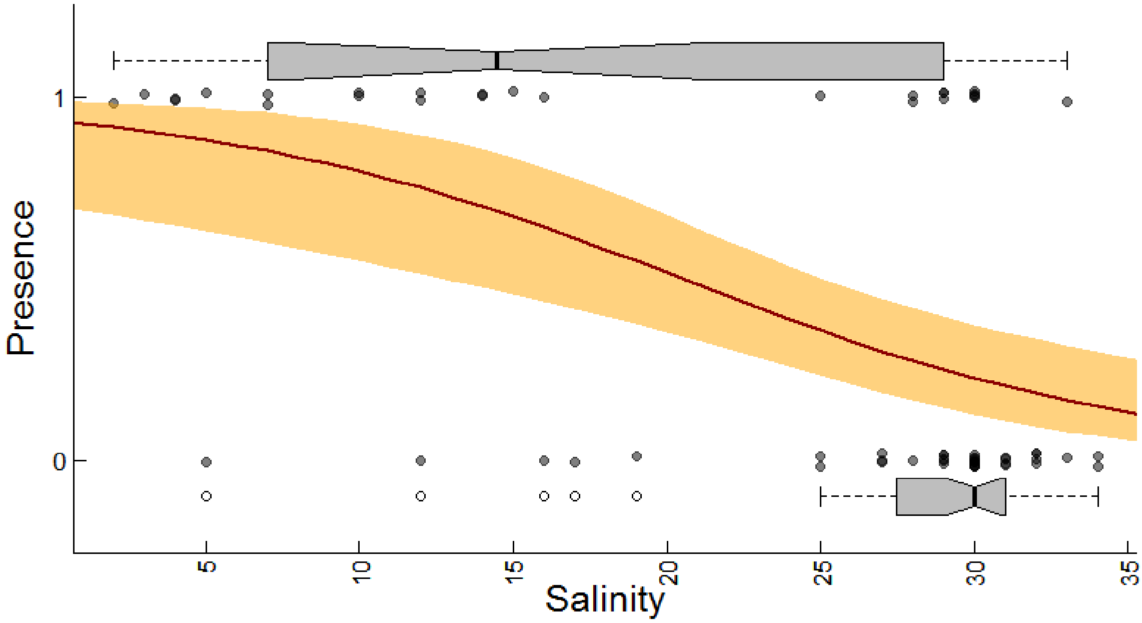
- `glm` - for **generalized** linear model (instead of `lm`)
- `family = 'binomial'` is the instruction to fit the logistic regression

```
glm(presence ~ salinity, family ='binomial')
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.661	0.902	2.951	0.003
salinity	-0.130	0.035	-3.716	0.000

Clearly - *Solea solea* presence is very significantly *negatively* related to salinity.

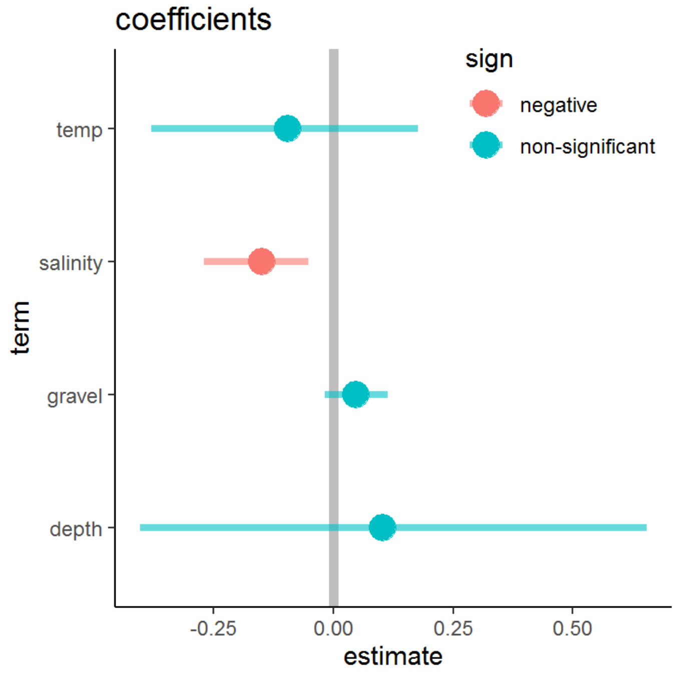
Out of this model we can make predictions



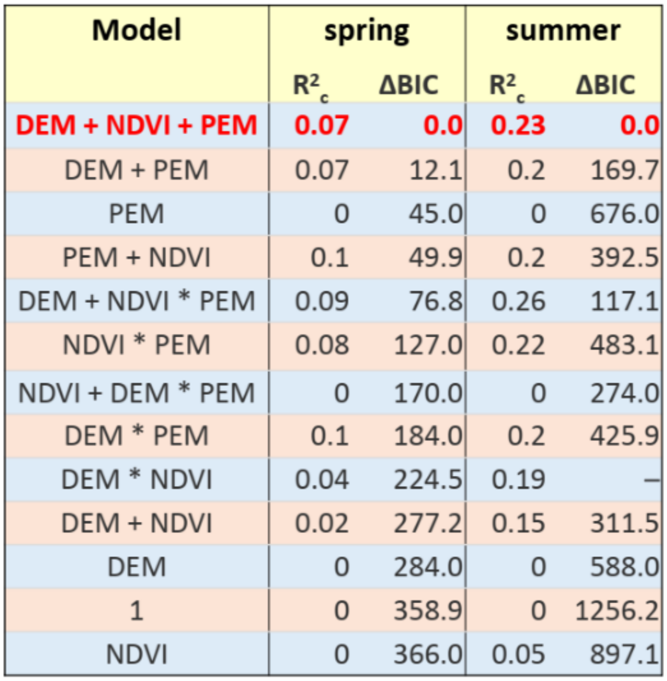
∆AIC analysis - and coefficients

	Model	k	logLik	AIC	dAIC
M9	**salinity + gravel**	**3**	**-33.2**	**72.5**	**0.0**
M2	**salinity**	**2**	**-34.3**	**72.6**	**0.1**
M7	**temp + salinity**	**3**	**-34.0**	**74.0**	**1.5**
M5	**depth + salinity**	**3**	**-34.1**	**74.3**	**1.8**
M11	depth + temp + salinity	4	-33.9	75.8	3.3
M0	depth	2	-38.1	80.1	7.6
M4	depth + temp	3	-38.0	81.9	9.4
M6	depth + gravel	3	-38.0	82.0	9.5
M10	depth + temp + gravel	4	-37.8	83.7	11.2
M1	temp	2	-43.3	90.6	18.1
M3	gravel	2	-43.7	91.3	18.8
M8	temp + gravel	3	-43.3	92.6	20.1

Salinity clearly among the more important covariates (in the top 4 models).



FLASHBAK: how the caribou Resource Selection Function was selected



Model	spring		summer	
	R^2_c	ΔBIC	R^2_c	ΔBIC
DEM + NDVI + PEM	**0.07**	**0.0**	**0.23**	**0.0**
DEM + PEM	0.07	12.1	0.2	169.7
PEM	0	45.0	0	676.0
PEM + NDVI	0.1	49.9	0.2	392.5
DEM + NDVI * PEM	0.09	76.8	0.26	117.1
NDVI * PEM	0.08	127.0	0.22	483.1
NDVI + DEM * PEM	0	170.0	0	274.0
DEM * PEM	0.1	184.0	0.2	425.9
DEM * NDVI	0.04	224.5	0.19	–
DEM + NDVI	0.02	277.2	0.15	311.5
DEM	0	284.0	0	588.0
1	0	358.9	0	1256.2
NDVI	0	366.0	0.05	897.1

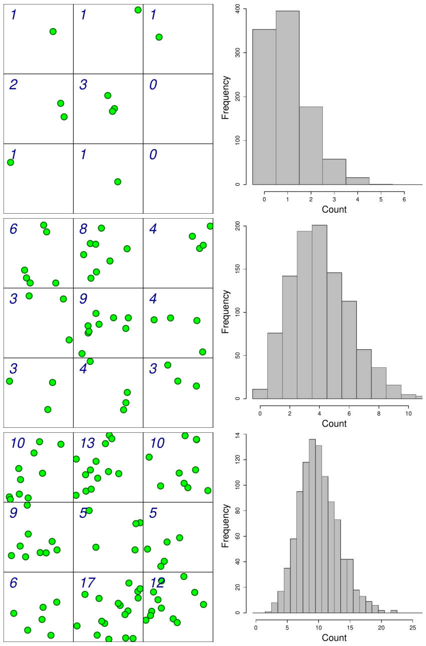
← ***THIS IS THE BEST MODEL!***
We will talk about why later.

Takeaways:

- For **both seasons** all **THREE** variables are important as main effects.
- **Summer** model explains *much more* (23%) than **Spring** model (7%).

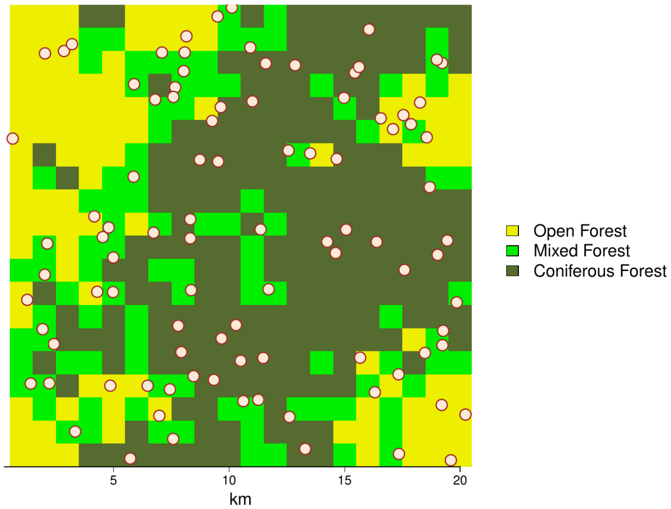
Note: "DEM" is second-order polynomial: DEM + DEM^2

Poisson regression



$$Y_i \sim \mathcal{Poisson}\left(\lambda = \exp(\alpha + \beta X_i)\right)$$

- We are **counting** something ... the data are between 0 and ∞
- λ is a **density**; **densities** vary across habitat types (covariate **X**).

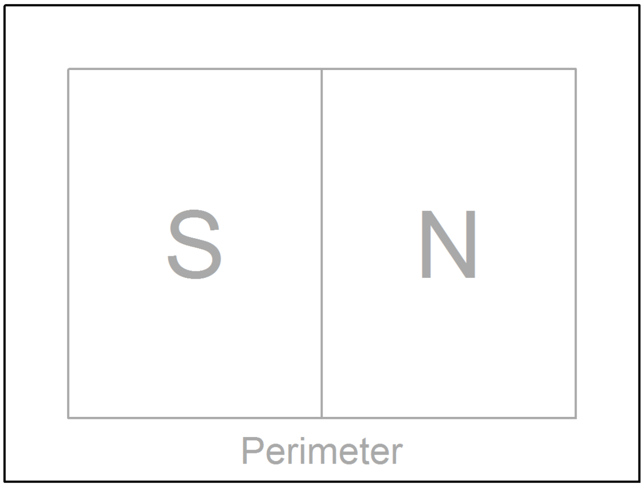


Field flags

Did flag densities vary with region?

Approximate areas:

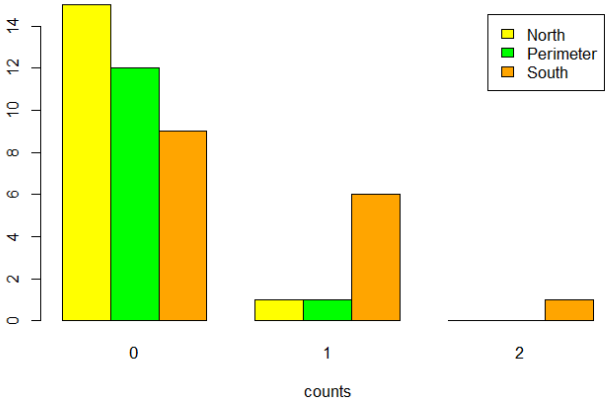
region	area
North:	82 m^2
South:	82 m^2
Perimeter:	196 m^2
--	--
Sampling square (hula hoop)	0.5 m^2



Count data

Lots of 0's, some 1's, and just one 2 count.

```
##       Region
## Count North Perimeter South
##     0    15        12     9
##     1     1         1     6
##     2     0         0     1
```



Fitting models

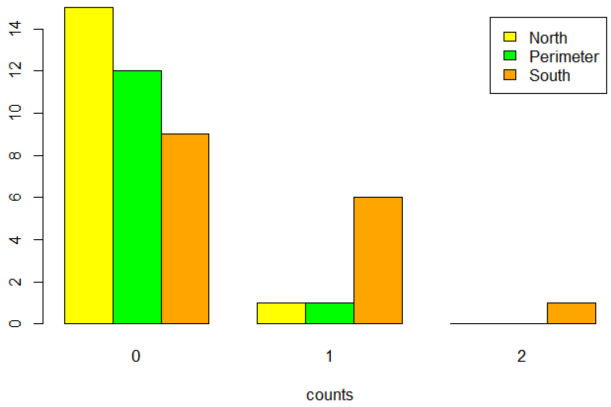
```
glm(count ~ region, family = 'poisson')
```

Exact same syntax as before, except the "family" is **Poisson.**

Count data

Lots of 0's, some 1's, and just one 2 count.

```
##      Region
## Count North Perimeter South
##     0    15        12     9
##     1     1         1     6
##     2     0         0     1
```



Fitting models

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.773	1.000	-2.773	0.006
RegionPerimeter	0.208	1.414	0.147	0.883
RegionSouth	2.079	1.061	1.961	0.050

The **intercept** here is "North", the *p*-values compare with North. So **South** has - borderline - significantly more

ΔAIC **table**

	df	AIC
Null.model	1	53.47
Region.model	3	49.15

Model that includes **Region** has lower AIC

Making predictions

Region	area	fit	se.fit	l.hat	l.low	l.high	d.hat	d.low	d.high	N.hat	N.low	N.high
South	82	-0.693	0.354	0.500	0.247	1.014	1.000	0.493	2.028	82.0	40.4	166.3
North	82	-2.773	1.000	0.063	0.008	0.462	0.125	0.017	0.924	10.2	1.4	75.8
Perimeter	196	-2.565	1.000	0.077	0.010	0.568	0.154	0.021	1.137	30.2	4.1	222.9

- **fit** and **se.fit** are in the log scale, so they need to be transformed via exp to intensities λ.
- `l.hat` is the Poisson intensity λ of the sampling square (**hula hoop**), which we turn into an actual density by dividing by its area **0.5 m^2**.
- `d.hat` (and `d.low` and `d.high`) are the density estimates & confidence intervals, which we then turn into our numerical predictions by multiplying by area.

Total estimate

$$\widehat{N} = 122.4\,(95\%\ \text{C.I.}: 71.4 - 173.4)$$

pretty darned good! The true values were 92 total [58 S, 29 N, 13 perimeter]

Take-aways on (linear, statistical) modeling

1. **Linear modeling** separates **patterns** (the model) from "**randomness**" (unexplained variation).

2. We structure our models to have a **response variable** and one or more **predictors** or **covariates**.

3. Depending on the reponse variable, a different **family** is chosen:
   - if **continuous** and symmatric: **Normal** family
   - if two values (presence/absence, dead/alive): **Binomial** family
   - if count data: **Poisson** family.

4. An improtant task is **Model selection**, identifying which model is "best"
   - Best means *"explains the most variation without overfitting"*
   - Very common criterion is **AIC.**

5. Once a model is "selected", we can:
   - analyze the results by seeing the **effect sizes** (magnitude of coefficients, aka *slopes*) and **directions** (signs of coefficients)
   - make **inferential predictions** by "spreading" our model over a larger landscape.

6. **Well over 90% of habitat modeling is done this way!**